Before we start ...

*** A few things to take note

- For those that have not created your AWS account, you can do so now
- Note that you have to attach a payment method, you can just listen through if you are uncomfortable with it

Amazon Sagemaker

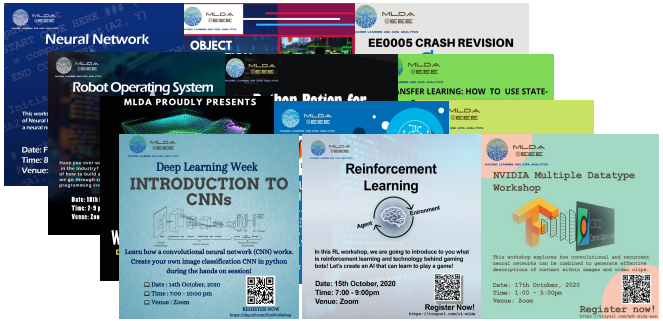
Philip Lee | 15 Sep 2022


MLDA@EEE
AY2022/23
MLDA @ EEE
MACHINE LEARNING AND DATA ANALYTICS

What we do

LEARN

Listen to our workshops



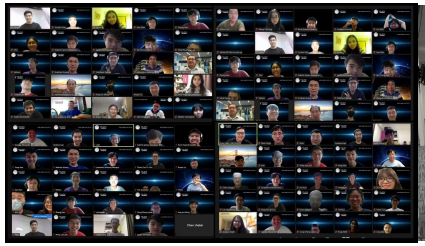
APPLY

Do projects that belong to yourselves

CONNECT

Hear from the industry

What we have



Like-minded Students in NTU

- Project Partner / Research Partner
- Friends in similar path of personal development

Academic facilities

- GPUs
  - RTX 3090, GTX 2080 Ti, GTX 1080 Ti
- Equipments
  - Robot Master *2
  - 3D printer *1
  - Drones *1
  - ...

Industrial Connections

- Sponsors & Partners

Follow us to stay tuned!

https://www.ntu.edu.sg/eee/student-life/mlda

Machine Learning and Data Analytics at EEE NTU

Machine Learning and Data Analytics Lab at NTU EEE

@mlda_at_eee_ntu

MLDA-NTU

TABLE OF CONTENTS

Before we start ...

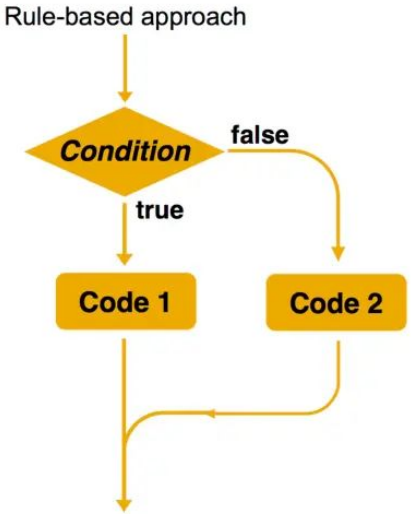
*** A few things to take note

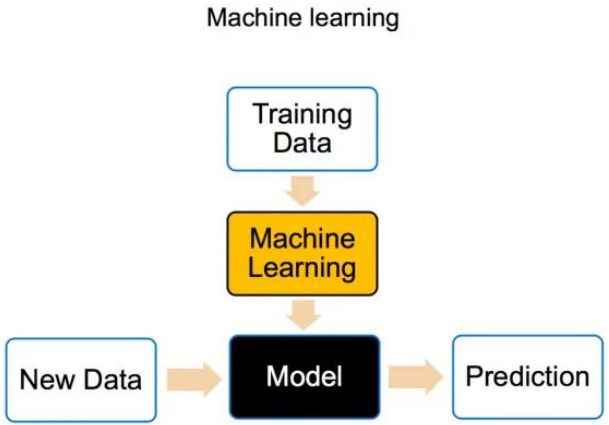
- Create AWS account
- Payment methods
  - You may just listen only
  - Session may cost 0-50¢
- Sharing session rather than a licensed AWS workshop
- Create SageMaker Studio domain
  - *us-west-1*

01

What is SageMaker?

What is machine learning?





Dog 99.99%!

What is cloud computing?

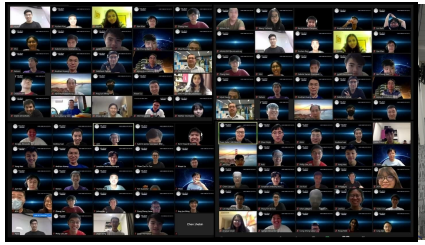
Access services on demand

Avoid large upfront investments

Provision computing resources as needed

Pay only for what you use

What we have



Like-minded Students in NTU

- Project Partner / Research Partner
- Friends in similar path of personal development

Academic facilities

- GPUs
  - RTX 3090, GTX 2080 Ti, GTX 1080 Ti
- Equipments
  - Robot Master *2
  - 3D printer *1
  - Drones *1
  - ...

Industrial Connections

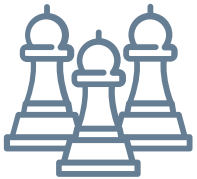
- Sponsors & Partners

Cloud Computing Benefits

Variable expenses

Cost optimization

Capacity

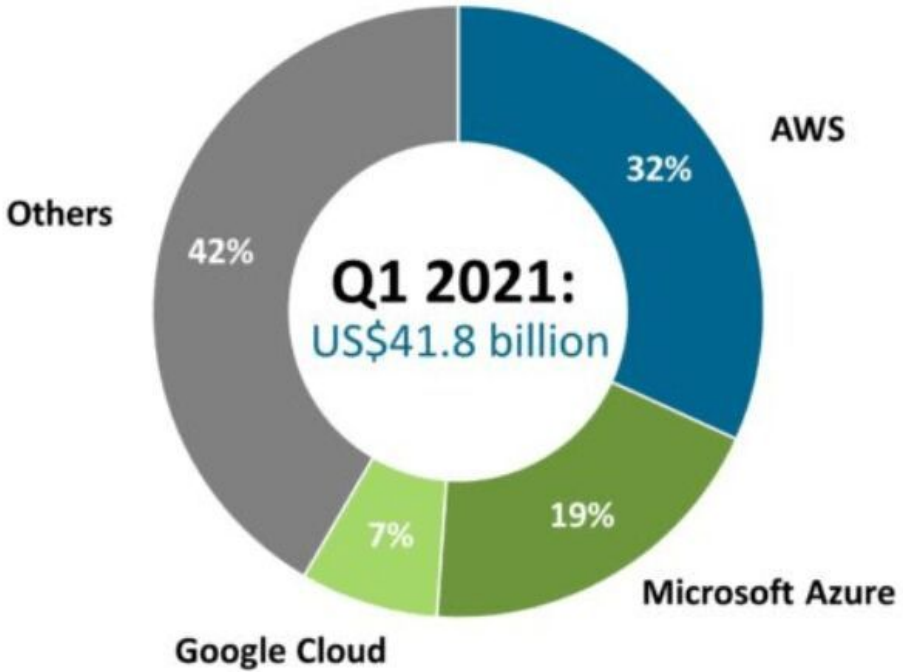


Economies of Scale

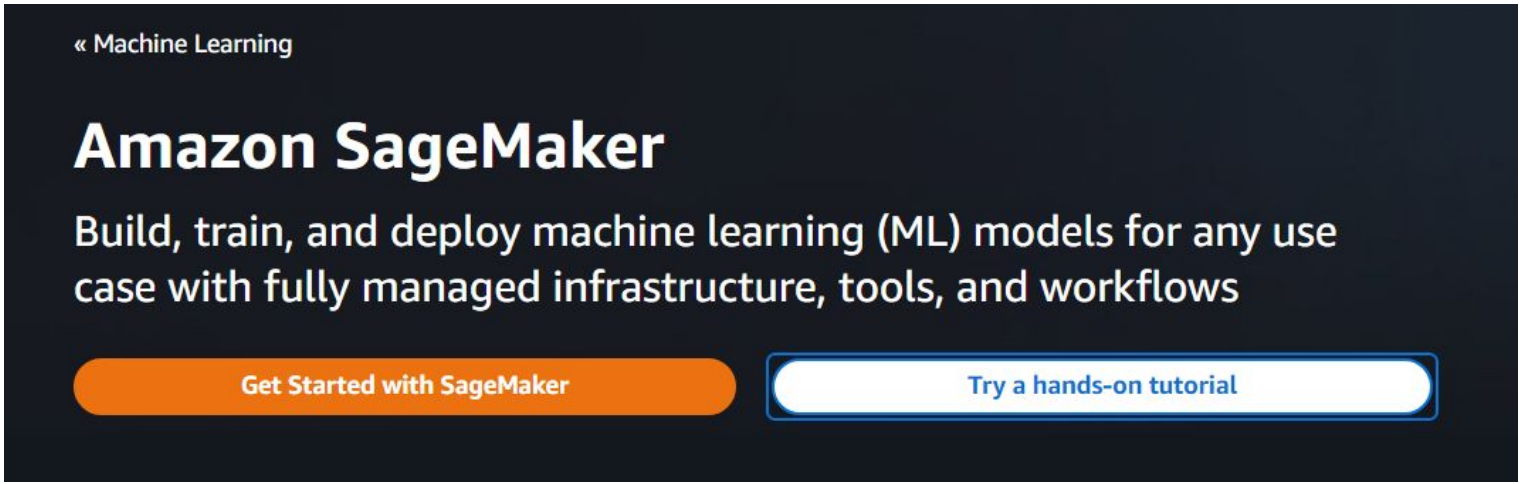
Speed and agility

Global in minutes

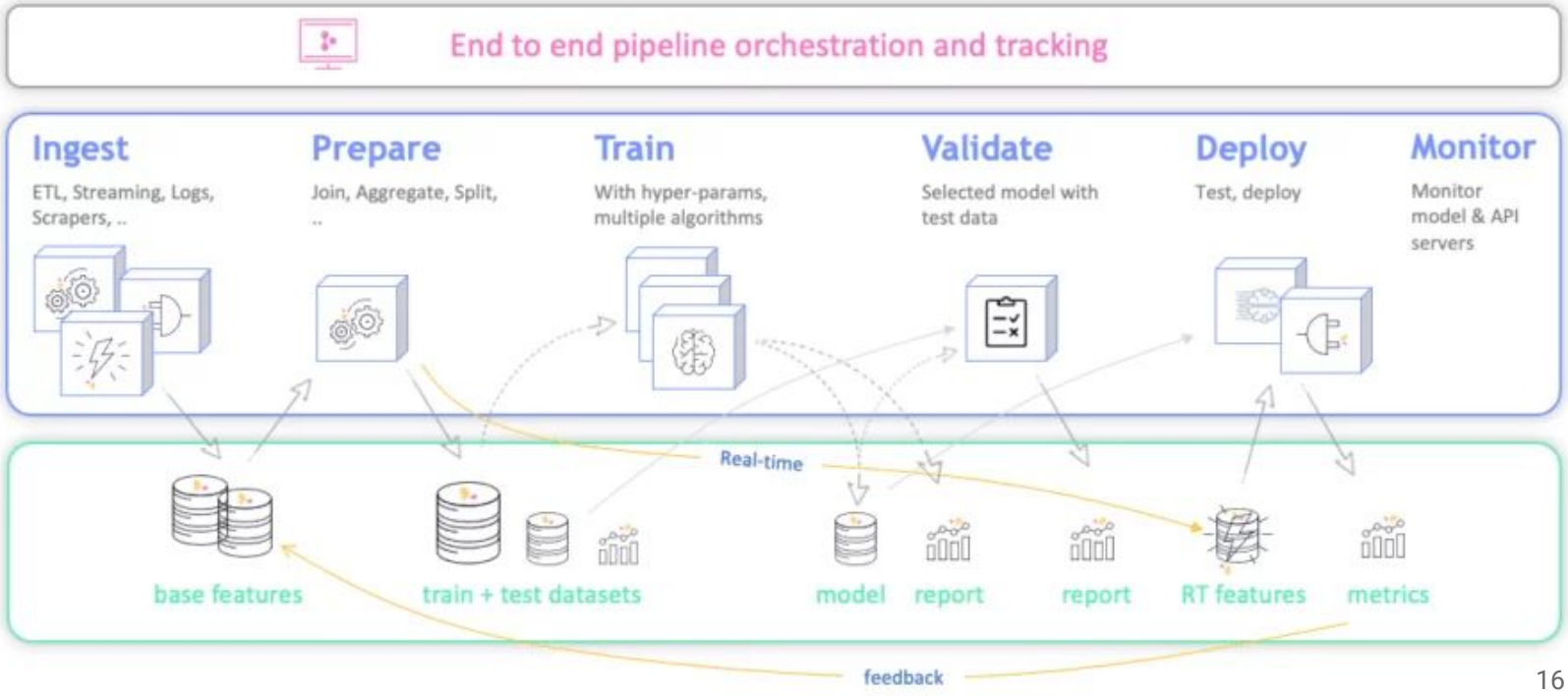
Major Cloud Providers



Where does SageMaker fit in?



End-to-end ML Pipeline



The AWS ML Stack

Broadest and most complete set of Machine Learning capabilities

AI SERVICES

VISION	SPEECH		TEXT			SEARCH	CHATBOTS	PERSONALIZATION	FORECASTING	FRAUD	DEVELOPMENT	CONTACT CENTERS
Amazon Rekognition	Amazon Polly	Amazon Transcribe *+Medical*	Amazon Comprehend *+Medical*	Amazon Translate	Amazon Textract	Amazon Kendra	Amazon Lex	Amazon Personalize	Amazon Forecast	Amazon Fraud Detector	Amazon CodeGuru	Contact Lens *For Amazon Connect*

ML SERVICES

Amazon SageMaker

Ground Truth

AWS Marketplace for ML

SageMaker Studio IDE

Built-in algorithms | Notebooks | Experiments | Processing | Model training & tuning | Debugger | Autopilot | Model hosting | Model Monitor

Neo

Augmented AI

ML FRAMEWORKS & INFRASTRUCTURE

TensorFlow | mxnet | PYTORCH

GLUON | Keras | DeepGraphLibrary

Deep Learning AMIs & Containers | GPUs & CPUs | Elastic Inference | Inferentia | FPGA

aws machine learning

Amazon SageMaker

Prepare

Ground Truth
Create high quality datasets for ML

Data Wrangler
Aggregate and prepare data for ML

Processing
Built-in Python, BYO R/Spark

Feature Store
Store, catalog, search, and reuse features

Clarify
Detect bias and understand model predictions

Build

Studio Notebooks & Notebook Instances
Fully managed Jupyter notebooks with elastic compute

Studio Lab
Free ML development environment

Built-in Algorithms
Integrated tabular, NLP, and vision algorithms

JumpStart
UI based discovery, training, and deployment of models, solutions, and examples

Autopilot
Automatically create ML models with full visibility

Bring Your Own
Bring your own container and algorithms

Local Mode
Test and prototype on your local machine

Train & tune

Fully Managed Training
Broad hardware options, easy to setup and scale

Distributed Training Libraries
High performance training for large datasets and models

Training Compiler
Faster deep learning model training

Automatic Model Tuning
Hyperparameter optimization

Managed Spot Training
Reduce training cost by up to 90%

Debugger and Profiler
Debug and profile training runs

Experiments
Track, visualize, and share model artifacts across teams

Customization Support
Integrate with popular open source frameworks and libraries

Deploy & manage

Fully Managed Deployment
Ultra low latency, high throughput inference

Real-Time Inference
For steady traffic patterns

Serverless Inference
For intermittent traffic patterns

Asynchronous Inference
For large payloads or long processing times

Batch Transform
For offline inference on batches of large datasets

Multi-Model Endpoints
Reduce cost by hosting multiple models per instance

Multi-Container Endpoints
Reduce cost by hosting multiple containers per instance

Inference Recommender
Automatically select compute instance and configuration

Model Monitor
Maintain accuracy of deployed models

Kubernetes & Kubeflow Integration
Simplify Kubernetes-based ML

Edge Manager
Manage and monitor models on edge devices

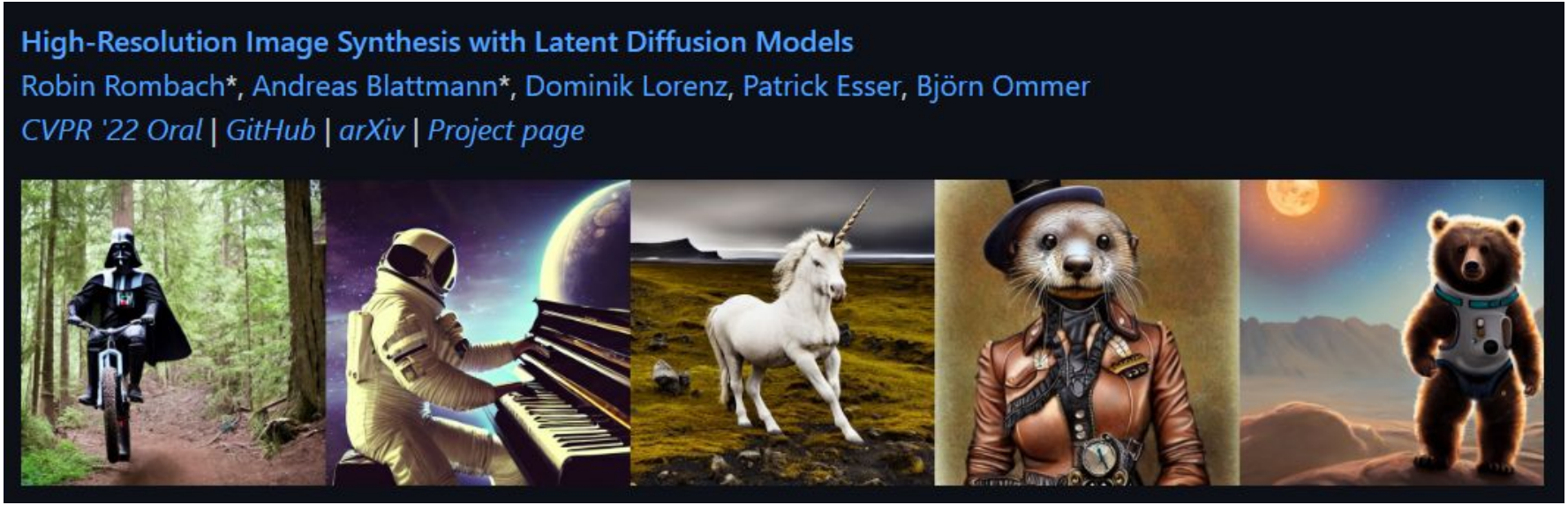
Studio | RStudio
Integrated development environment (IDE) for ML

MLOps: Pipelines | Projects | Model Registry
Workflow automation, CI/CD for ML, central model catalog

Canvas
Generate accurate machine learning predictions—no code required

Stability AI: Stable Diffusion

Stability AI: Stable Diffusion - Open Sourced



https://github.com/CompVis/stable-diffusion https://www.youtube.com/watch?v=nVhmFski3vg

Stability AI: Stable Diffusion - Model Card

Environmental Impact

Stable Diffusion v1 Estimated Emissions Based on that information, we estimate the following CO2 emissions using the Machine Learning Impact calculator presented in Lacoste et al. (2019). The hardware, runtime, cloud provider, and compute region were utilized to estimate the carbon impact.

- **Hardware Type:** A100 PCIe 40GB
- **Hours used:** 150000
- **Cloud Provider:** AWS
- **Compute Region:** US-east
- **Carbon Emitted (Power consumption x Time x Carbon produced based on location of power grid):** 11250 kg CO2 eq.

02

SageMaker Pricing

Free Tier – https://aws.amazon.com/sagemaker/pricing/

Amazon SageMaker capability	Free Tier usage per month for the first 2 months
Studio notebooks, and notebook instances	250 hours of ml.t3.medium instance on Studio notebooks OR 250 hours of ml.t2 medium instance or ml.t3.medium instance on notebook instances
RStudio on SageMaker	250 hours of ml.t3.medium instance on RSession app AND free ml.t3.medium instance for RStudioServerPro app
Data Wrangler	25 hours of ml.m5.4xlarge instance
Feature Store	10 million write units, 10 million read units, 25 GB storage
Training	50 hours of m4.xlarge or m5.xlarge instances
Real-Time Inference	125 hours of m4.xlarge or m5.xlarge instances
Serverless Inference	150,000 seconds of inference duration
Canvas	750 hours/month for session time, and up to 10 model creation requests/month, each with up to 1 million cells/model creation request

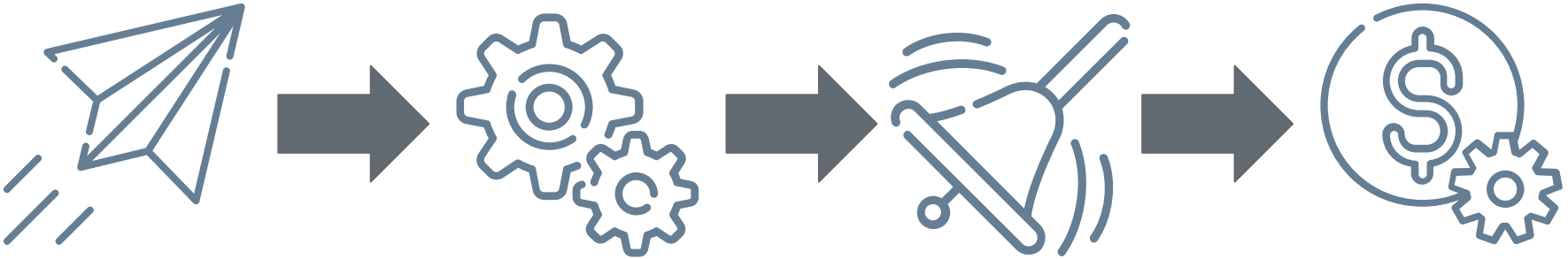
03

AWS Products

AWS Lambda - Serverless Computing

- Run code without provisioning or managing servers
- Pay only for compute time while code is running
- Use other AWS services to automatically trigger code

How AWS Lambda works



Upload code to Lambda

Set code to trigger from an event source

Code runs only when triggered

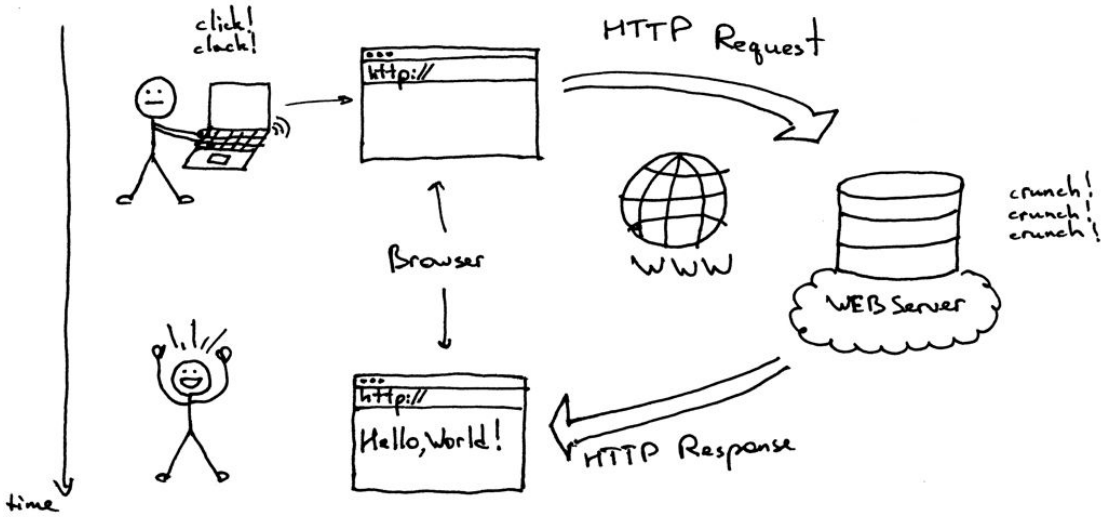
Pay only for the compute time you use

API Gateway

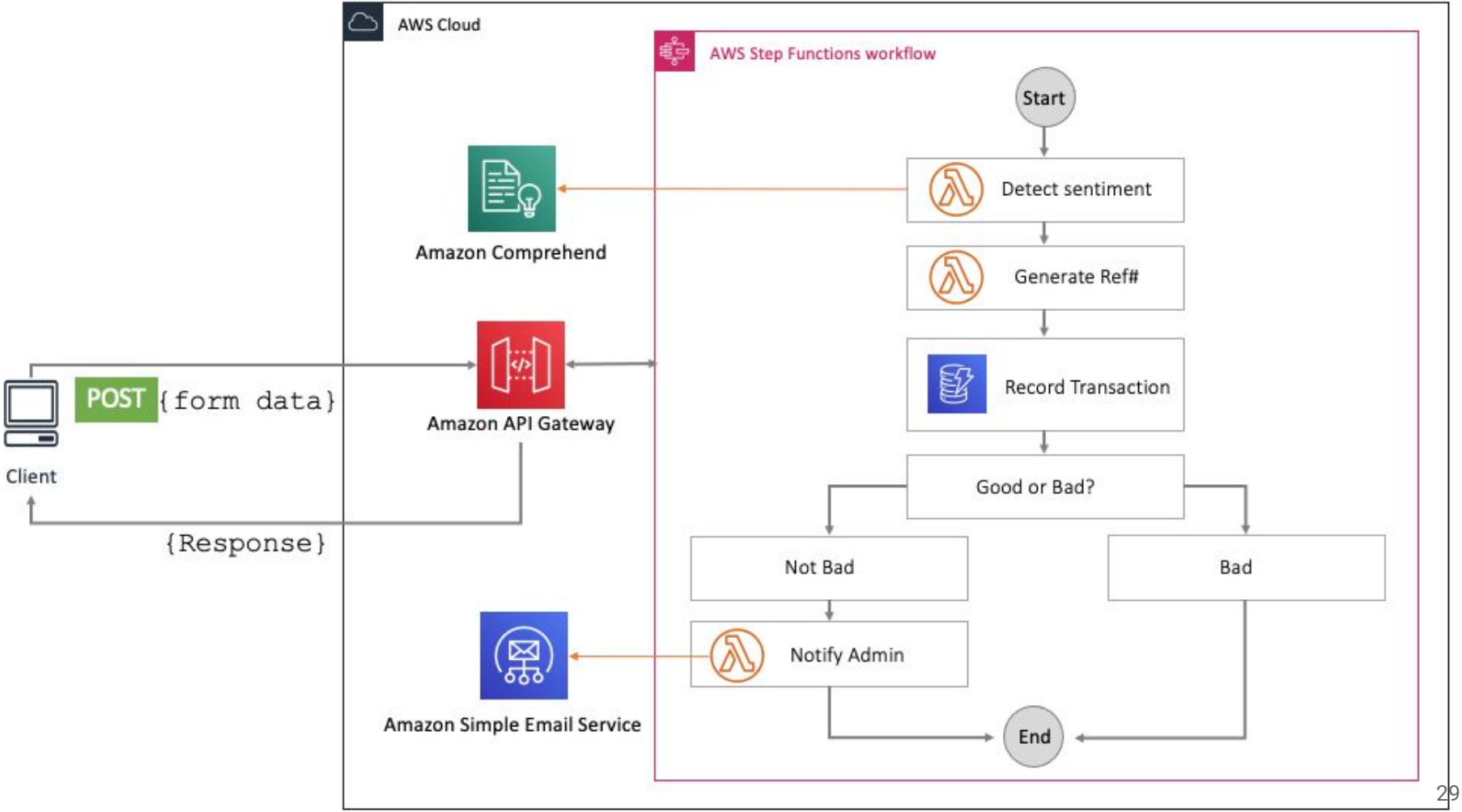
Acts as the "front door" for applications to access data, business logic, or functionality from your backend services.

Handles all the tasks involving:
- including traffic management
- CORS support
- authorization and access control
- throttling and monitoring
- API version management

HTTP Methods



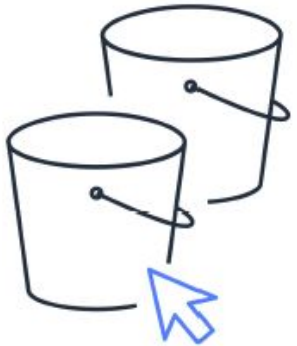
Method	CRUD Operation	Meaning
GET	Read	Read data
POST	Create	Insert data
PATCH	Update	Update data
DELETE	Delete	Delete data


AWS Cloud
AWS Step Functions workflow
Start
Amazon Comprehend
Detect sentiment
Generate Ref#
Record Transaction
POST {form data}
Amazon API Gateway
Client
Good or Bad?
{Response}
Not Bad
Bad
Notify Admin
Amazon Simple Email Service
End

Elastic Container Registry (ECR)

Push container images to Amazon ECR without installing or scaling infrastructure, and pull images using any management tool.

Amazon Simple Storage Service (S3)



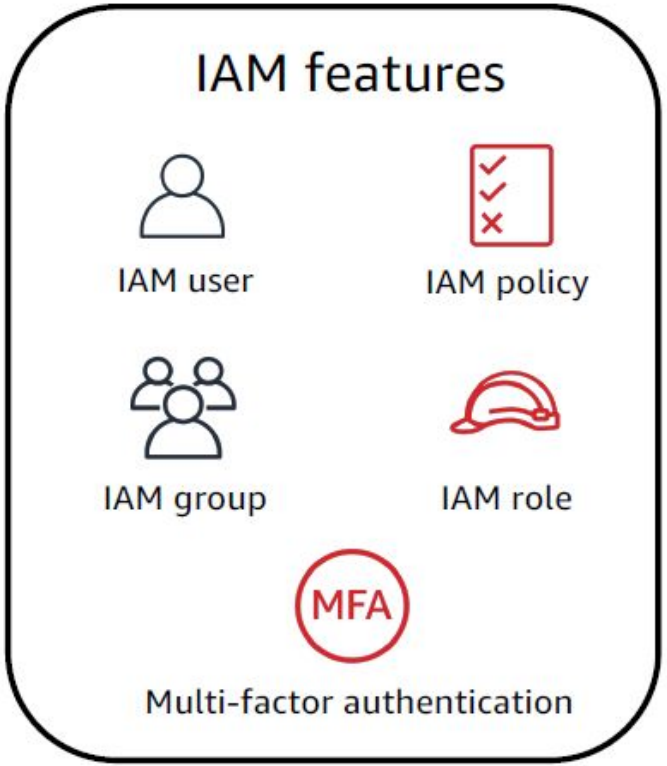
Store objects in buckets

Set permissions to control access to objects

Choose from a range of storage classes for different use cases

IAM

AWS Identity and Access Management (IAM) allows you to manage access to AWS services and resources.



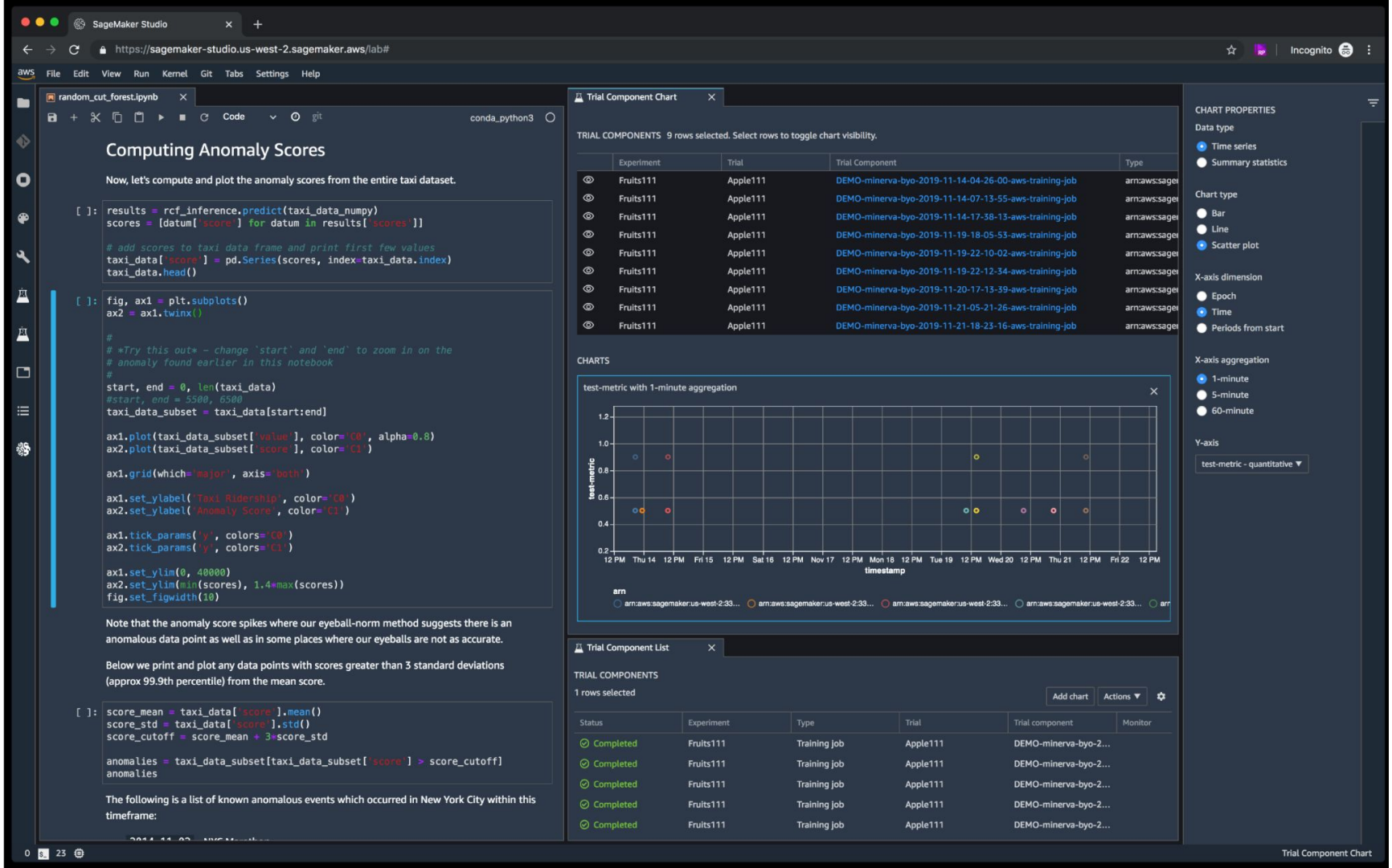
IAM policy

An IAM policy is a document that grants or denies permissions to AWS services and resources.

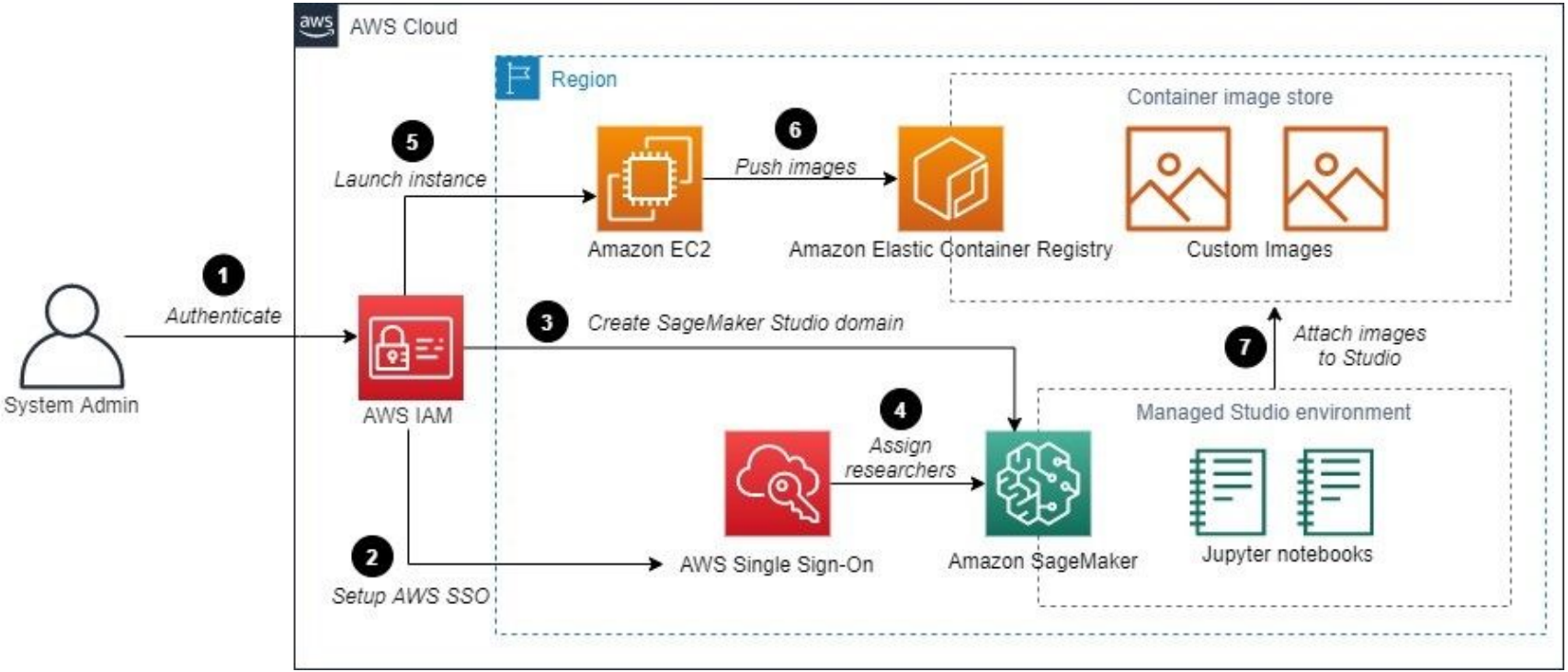
Best practice: Follow the security principle of least privilege.

04

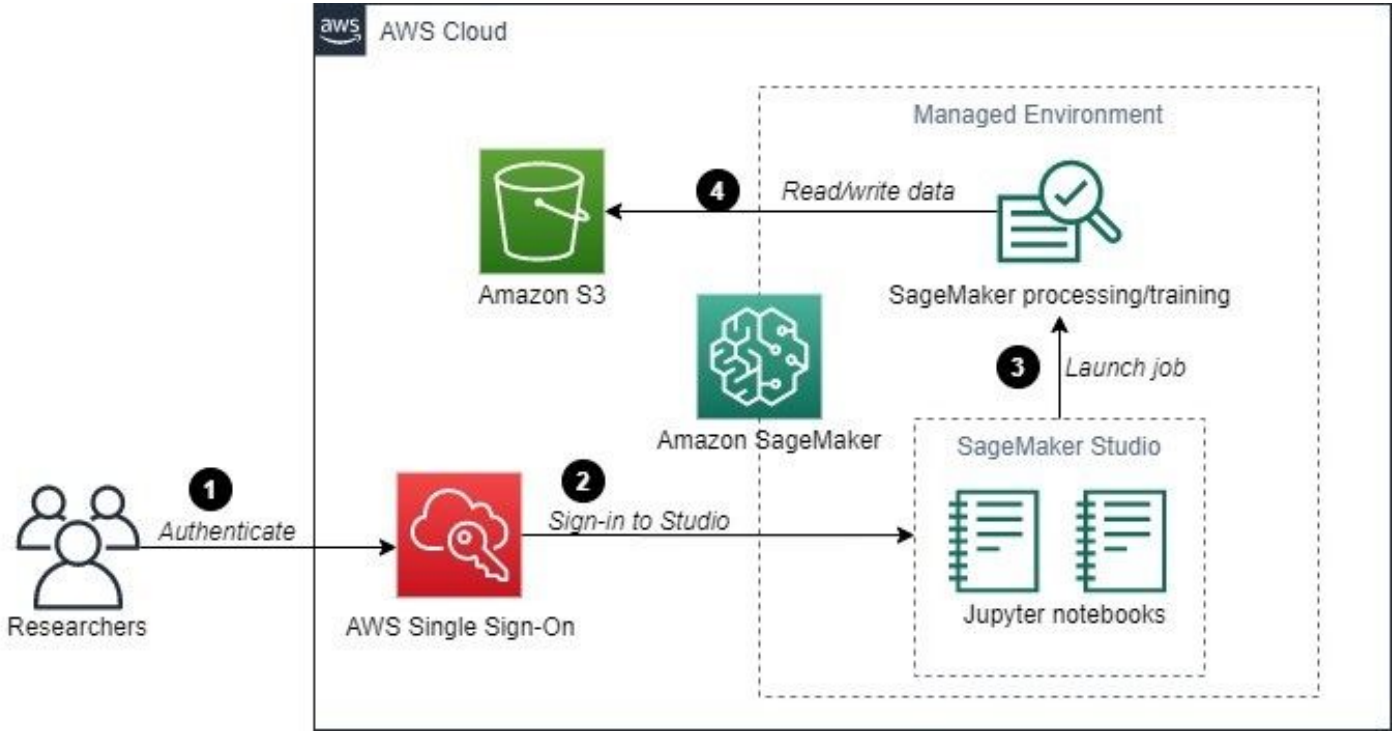
SageMaker Studio


SageMaker Studio
https://sagemaker-studio.us-west-2.sagemaker.aws/lab#
Incognito
File Edit View Run Kernel Git Tabs Settings Help
random_cut_forest.ipynb
Code
git
conda_python3
Computing Anomaly Scores
Now, let's compute and plot the anomaly scores from the entire taxi dataset.
[ ]:
results = rcf_inference.predict(taxi_data_numpy)
scores = [datum['score'] for datum in results['scores']]
# add scores to taxi data frame and print first few values
taxi_data['score'] = pd.Series(scores, index=taxi_data.index)
taxi_data.head()
[ ]:
fig, ax1 = plt.subplots()
ax2 = ax1.twinx()
#
Try this out - change `start` and `end` to zoom in on the
anomaly found earlier in this notebook
#
start, end = 0, len(taxi_data)
#start, end = 5500, 6500
taxi_data_subset = taxi_data[start:end]
ax1.plot(taxi_data_subset['value'], color='C0', alpha=0.8)
ax2.plot(taxi_data_subset['score'], color='C1')
ax1.grid(which='major', axis='both')
ax1.set_ylabel('Taxi Ridership', color='C0')
ax2.set_ylabel('Anomaly Score', color='C1')
ax1.tick_params('y', colors='C0')
ax2.tick_params('y', colors='C1')
ax1.set_ylim(0, 40000)
ax2.set_ylim(min(scores), 1.4*max(scores))
fig.set_figwidth(10)
Note that the anomaly score spikes where our eyeball-norm method suggests there is an anomalous data point as well as in some places where our eyeballs are not as accurate.
Below we print and plot any data points with scores greater than 3 standard deviations (approx 99.9th percentile) from the mean score.
[ ]:
score_mean = taxi_data['score'].mean()
score_std = taxi_data['score'].std()
score_cutoff = score_mean + 3*score_std
anomalies = taxi_data_subset[taxi_data_subset['score'] > score_cutoff]
anomalies
The following is a list of known anomalous events which occurred in New York City within this timeframe:
Trial Component Chart
TRIAL COMPONENTS 9 rows selected. Select rows to toggle chart visibility.
Experiment Trial Trial Component Type
Fruits111 Apple111 DEMO-minerva-byo-2019-11-14-04-26-00-aws-training-job arn:aws:sage
Fruits111 Apple111 DEMO-minerva-byo-2019-11-14-07-13-55-aws-training-job arn:aws:sage
Fruits111 Apple111 DEMO-minerva-byo-2019-11-14-17-38-13-aws-training-job arn:aws:sage
Fruits111 Apple111 DEMO-minerva-byo-2019-11-19-18-05-53-aws-training-job arn:aws:sage
Fruits111 Apple111 DEMO-minerva-byo-2019-11-19-22-10-02-aws-training-job arn:aws:sage
Fruits111 Apple111 DEMO-minerva-byo-2019-11-19-22-12-34-aws-training-job arn:aws:sage
Fruits111 Apple111 DEMO-minerva-byo-2019-11-20-17-13-39-aws-training-job arn:aws:sage
Fruits111 Apple111 DEMO-minerva-byo-2019-11-21-05-21-26-aws-training-job arn:aws:sage
Fruits111 Apple111 DEMO-minerva-byo-2019-11-21-18-23-16-aws-training-job arn:aws:sage
CHARTS
test-metric with 1-minute aggregation
test-metric
timestamp
12 PM Thu 14 12 PM Fri 15 12 PM Sat 16 12 PM Nov 17 12 PM Mon 18 12 PM Tue 19 12 PM Wed 20 12 PM Thu 21 12 PM Fri 22 12 PM
arn
arn:aws:sagemaker:us-west-2:33...
Trial Component List
TRIAL COMPONENTS
1 rows selected
Add chart
Actions
Status Experiment Type Trial Trial component Monitor
Completed Fruits111 Training job Apple111 DEMO-minerva-byo-2...
CHART PROPERTIES
Data type
Time series
Summary statistics
Chart type
Bar
Line
Scatter plot
X-axis dimension
Epoch
Time
Periods from start
X-axis aggregation
1-minute
5-minute
60-minute
Y-axis
test-metric - quantitative
Trial Component Chart

Admin Perspective



User Perspective

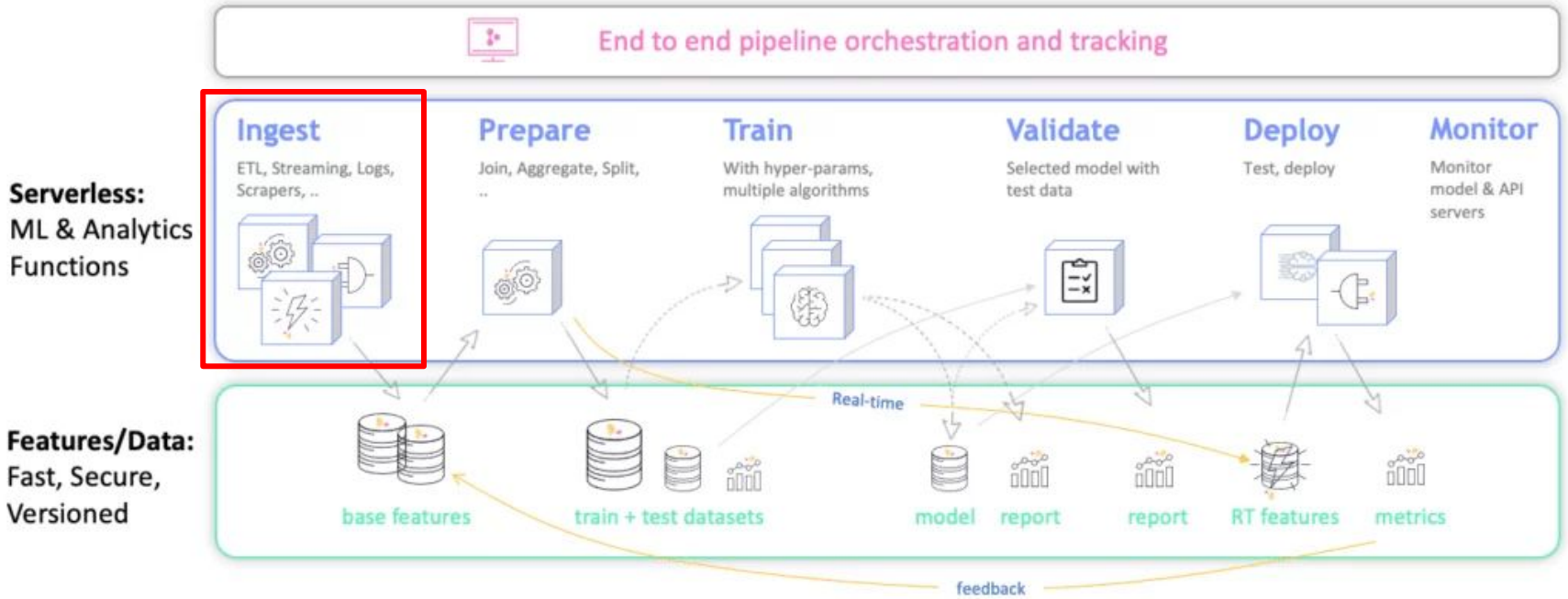


05

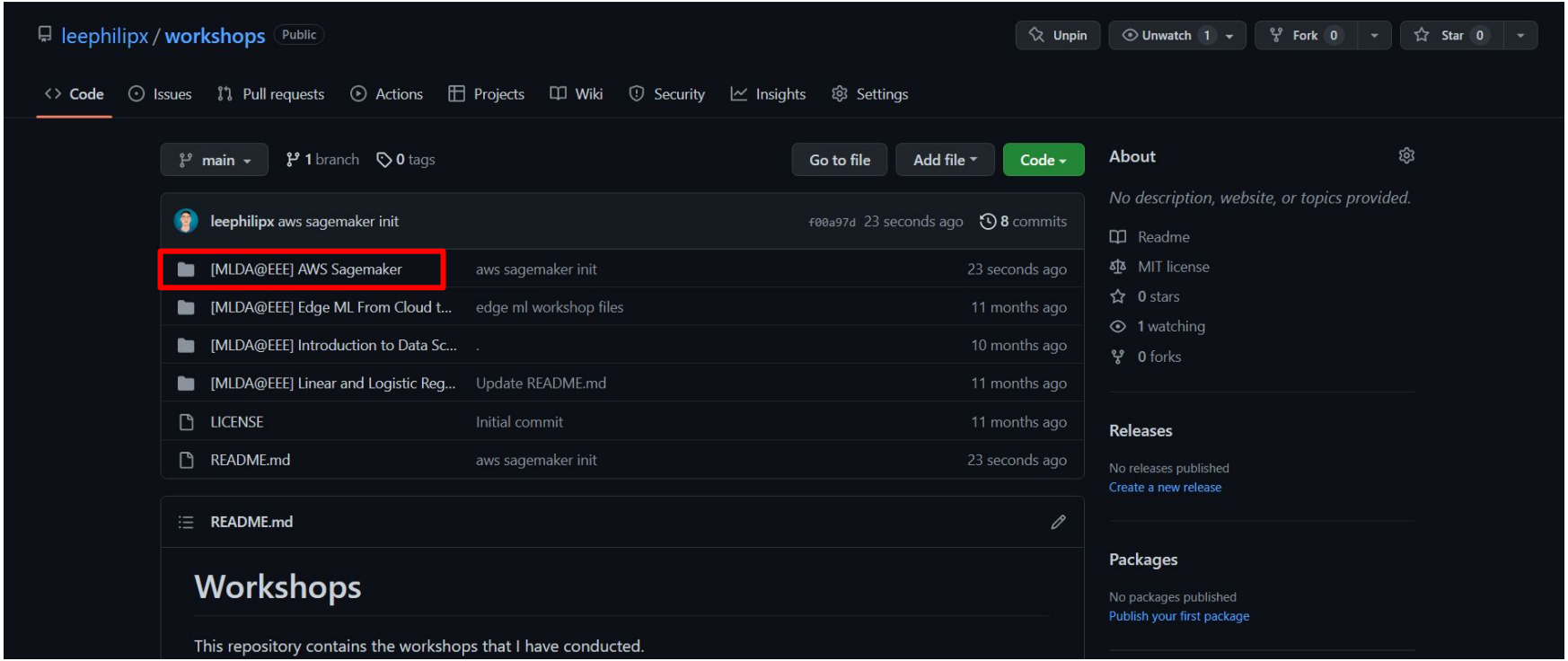
Demo + Hands-on

Demo: Labelling Data

https://aws.amazon.com/getting-started/hands-on/machine-learning-tutorial-label-training-data/



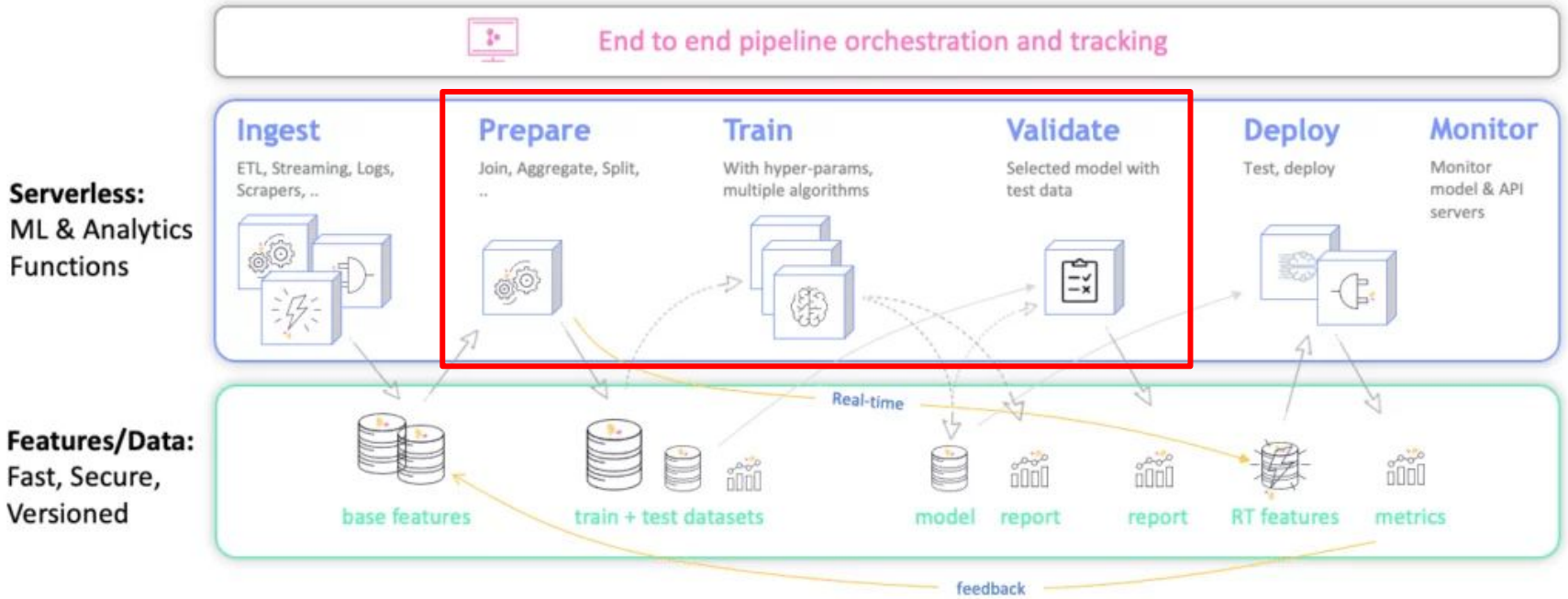
Code Resource – https://github.com/leephilipx/workshops



Hands-on Session

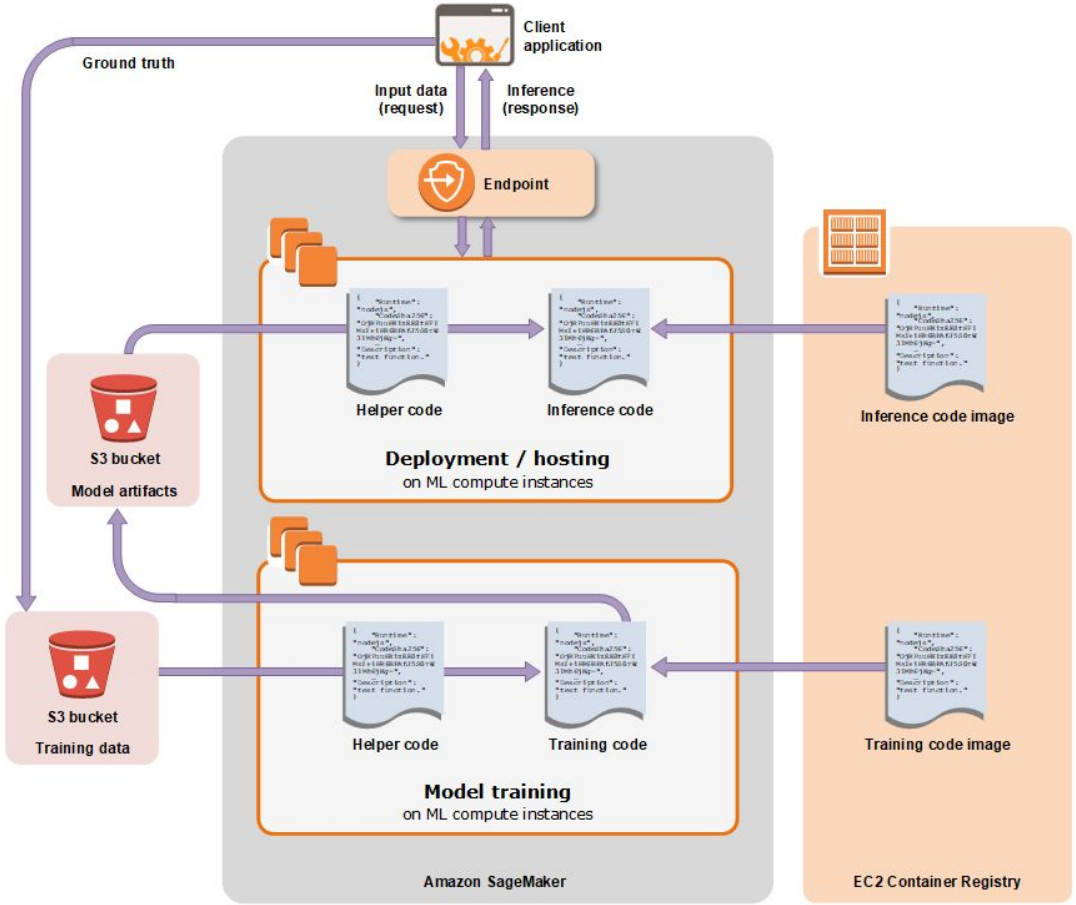
Hands-on: Training the Model

https://aws.amazon.com/getting-started/hands-on/machine-learning-tutorial-label-training-data/



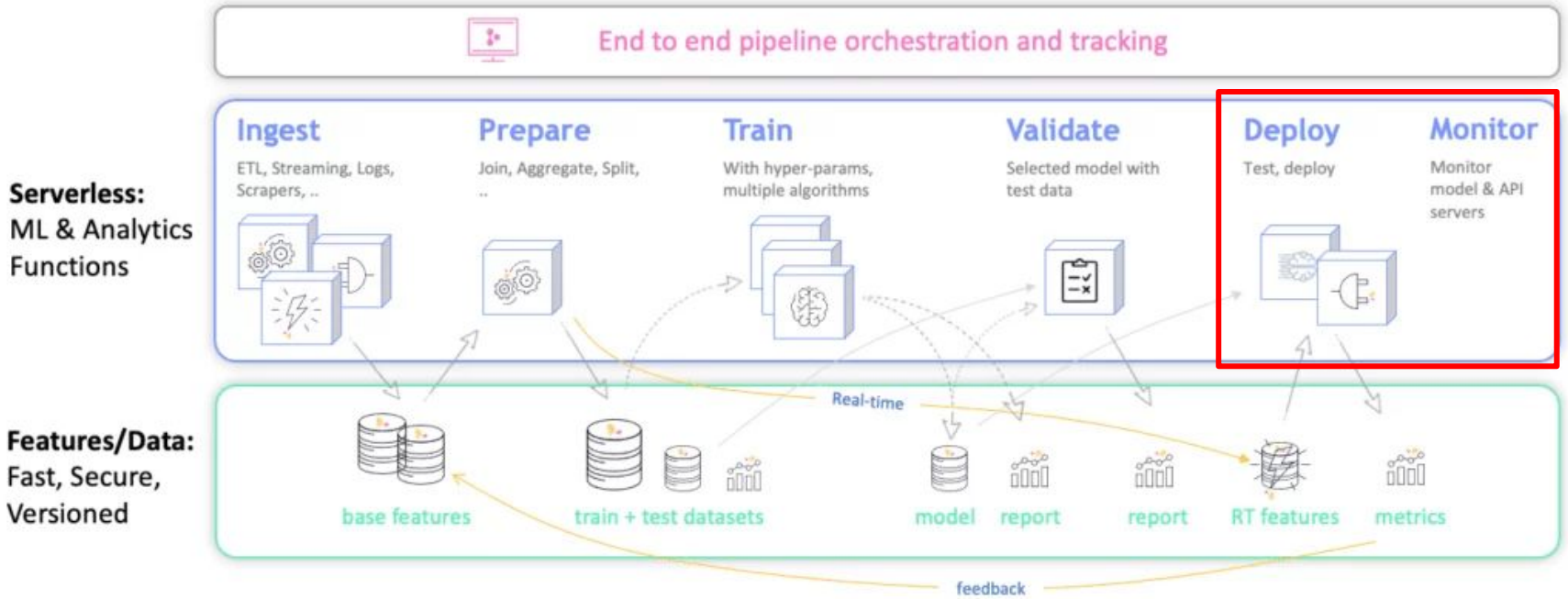
Training a Model

Under the Hood

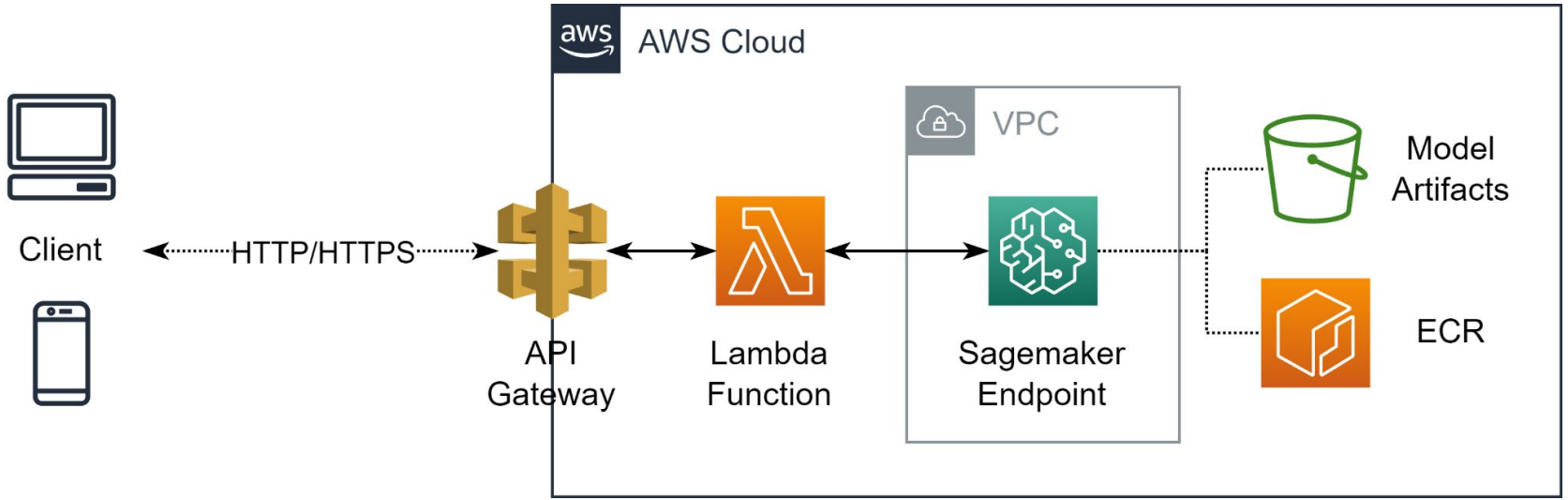


Hands-on: Endpoint Deployment

https://aws.amazon.com/getting-started/hands-on/machine-learning-tutorial-label-training-data/



Hands-on: Training & Deploying a Model

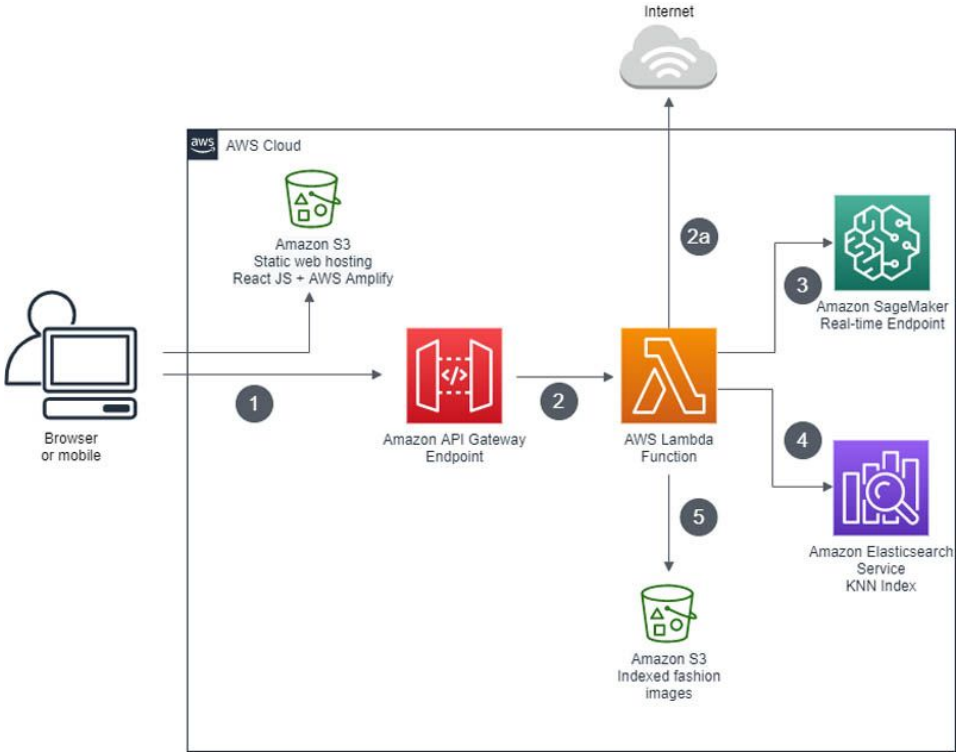


Hands-on Session

06

Sample Architectures

Visual Search Application



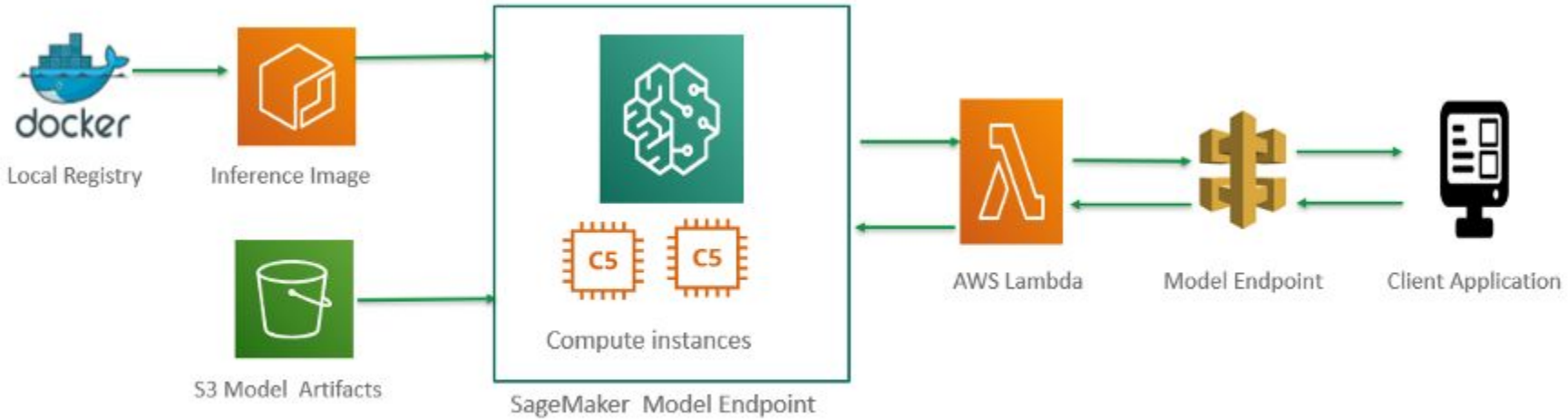
https://aws.amazon.com/blogs/machine-learning/building-a-visual-search-application-with-amazon-sagemaker-and-amazon-es/

API Gateway: Extra Setup

- API Authentication
- Rate-limiting
- Resource monitoring

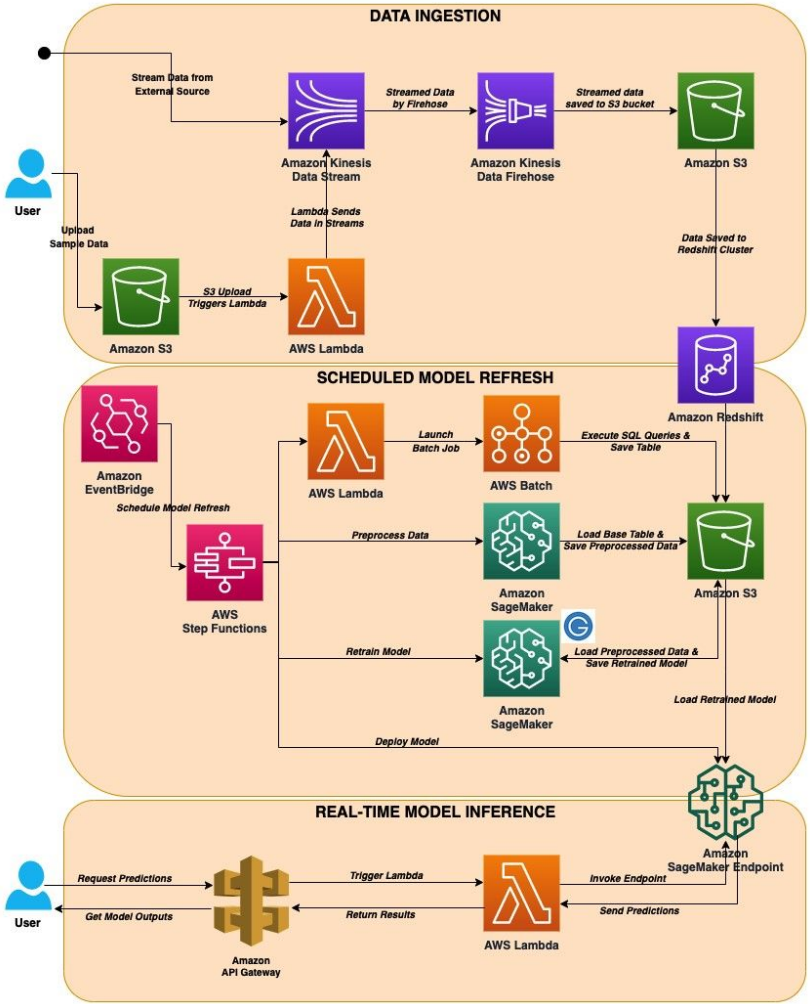
Already have a Docker image?

https://medium.com/@info_91596/deploying-a-custom-machine-learning-model-as-rest-api-with-aws-sagemaker-7216ba600504

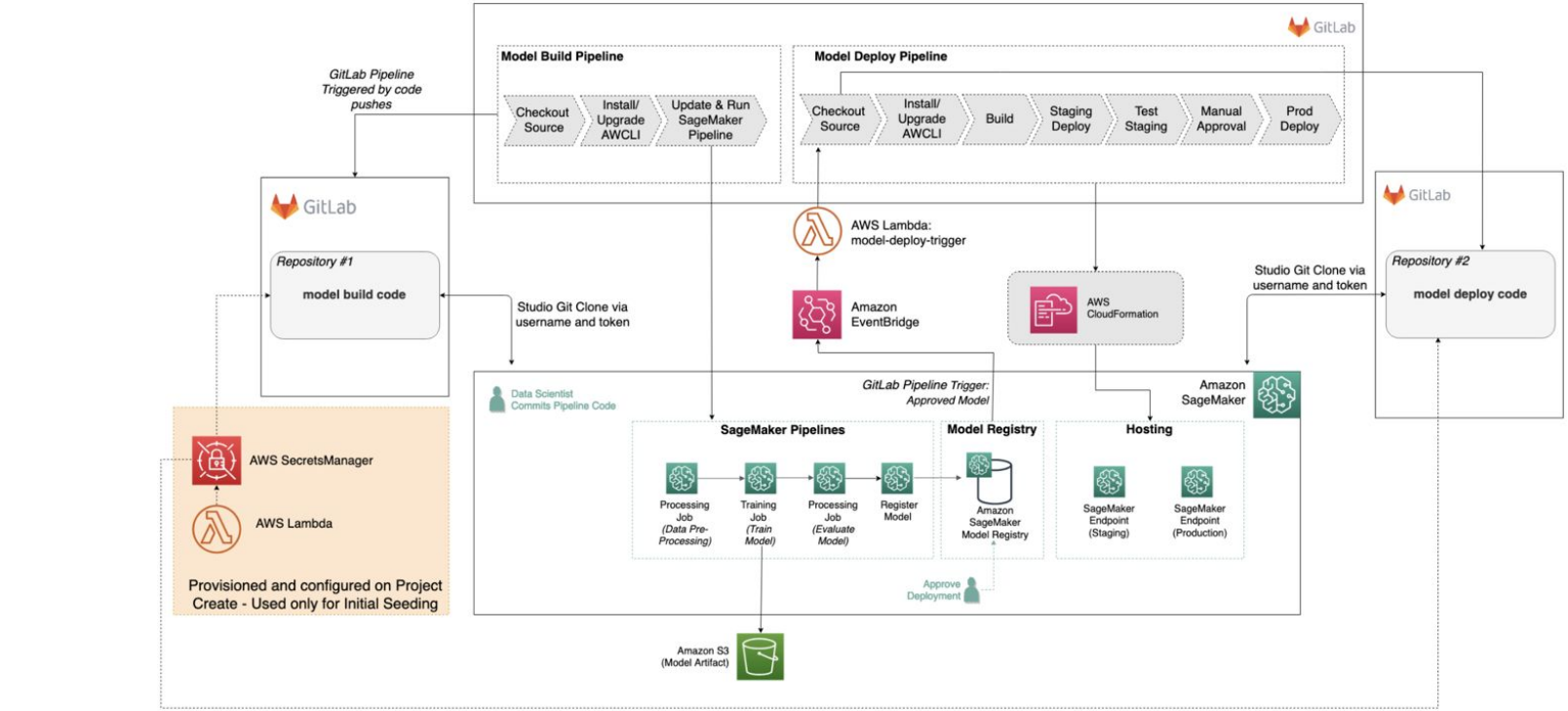


Automated model refresh with streaming data

https://aws.amazon.com/blogs/machine-learning/automated-model-refresh-with-streaming-data/



MLOps - CI/CD Pipeline



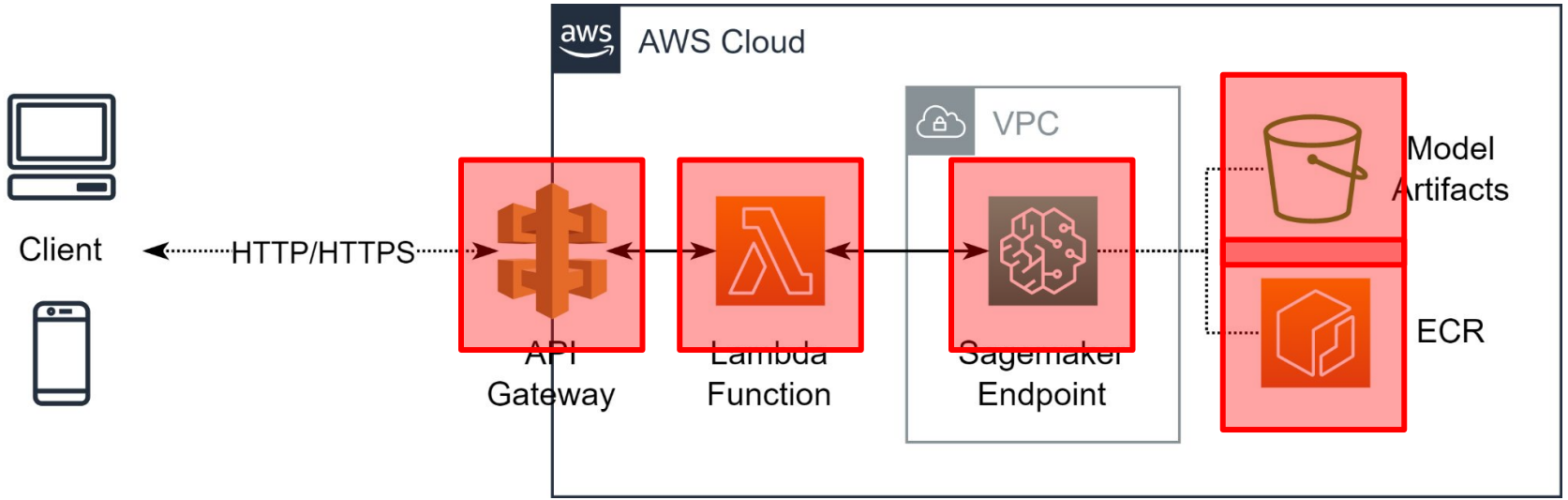
https://aws.amazon.com/blogs/machine-learning/build-mlops-workflows-with-amazon-sagemaker-projects-gitlab-and-gitlab-pipelines/

07

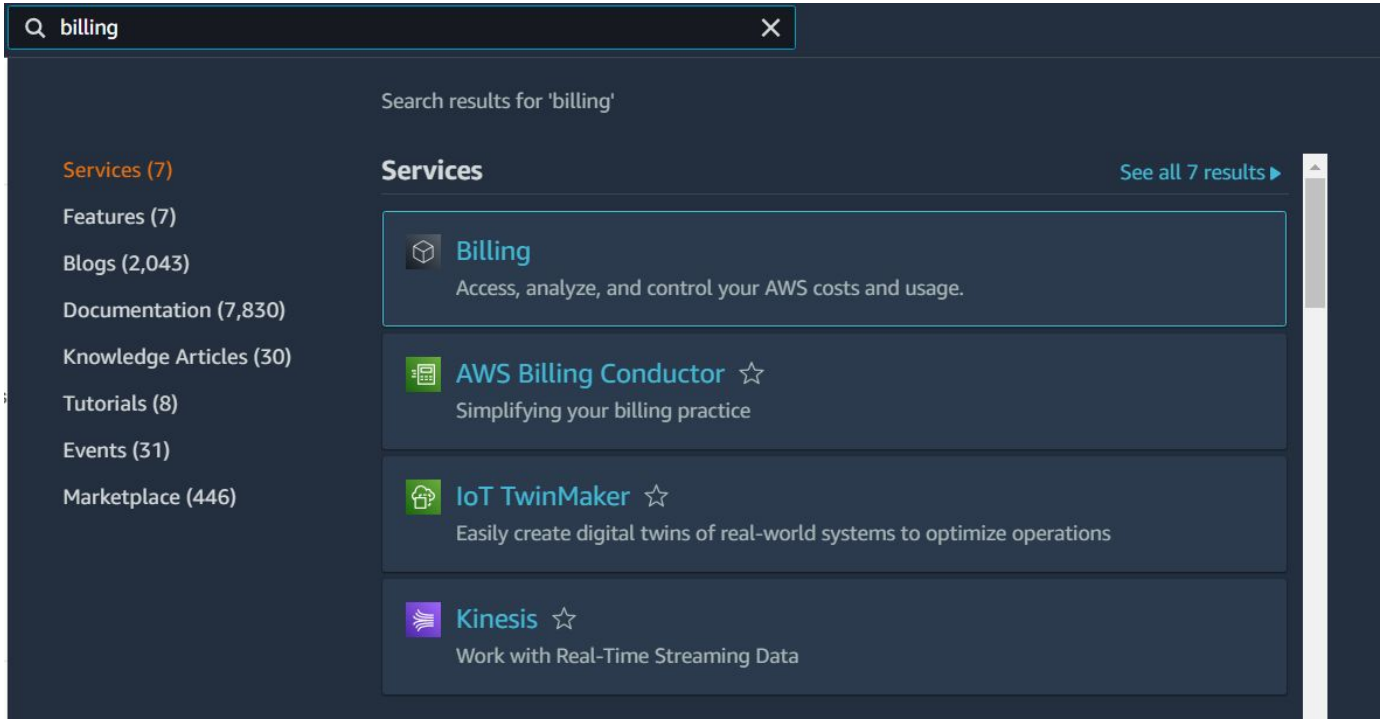
Cleaning-up

Delete your resources, save your wallet

Remember to free up all resources to avoid shocking bills at the end of the month!



Billings and Budgets



THE END

Thanks for attending!

Please leave your feedback

https://forms.office.com/r/zbUEJjEUHJ

Extra Resources

- AWS Official Documentation
- AWS Training
- AWS Certification

- SageMaker Developer Guide
- SageMaker Getting Started
- Amazon SageMaker Tutorials
- Amazon SageMaker 101 Workshop